

# Fast and Robust Detection of Fallen People from a Mobile Robot

Morris Antonello, Marco Carraro, Marco Pierobon and Emanuele Menegatti

**Abstract**—This paper deals with the problem of detecting fallen people lying on the floor by means of a mobile robot equipped with a 3D depth sensor. In the proposed algorithm, inspired by semantic segmentation techniques, the 3D scene is over-segmented into small patches. Fallen people are then detected by means of two SVM classifiers: the first one labels each patch, while the second one captures the spatial relations between them. This novel approach showed to be robust and fast. Indeed, thanks to the use of small patches, fallen people in real cluttered scenes with objects side by side are correctly detected. Moreover, the algorithm can be executed on a mobile robot fitted with a standard laptop making it possible to exploit the 2D environmental map built by the robot and the multiple points of view obtained during the robot navigation. Additionally, this algorithm is robust to illumination changes since it does not rely on RGB data but on depth data. All the methods have been thoroughly validated on the IASLAB-RGBD Fallen Person Dataset, which is published online as a further contribution. It consists of several static and dynamic sequences with 15 different people and 2 different environments.

## I. INTRODUCTION

In the richest countries, the population pyramid is turning upside down [1]. In 2015, 8.5 percent of the world's population was aged 65 and over and, by 2050, this older population is projected to represent 16.7 percent of the world total population. To allow people to continue to have active and productive lives as they age, new technologies are being studied. Recently, as far as home robots are concerned, there has been many promising developments. New products like Softbank's Pepper have been introduced into the market and many research platforms, e.g. the healthcare robots Pearl [2], ASTRO [3], Max [4], Hobbit [5] or our prototype O-Robot [6], have been proposed. Not only such robots aim at fostering research to keep the house safe by monitoring and detecting anomalies, but also at being friendly companions able to enhance the elderly people's social lives without invading their privacy. In particular, among all the sources of harm, falls are known to be the major one in elderly people [7]. In this work, given that it is unlikely for a robot to capture the act of falling while patrolling, the focus is on detecting people already lying on the floor.

The main contributions in this paper are:

- a real-time pure-3D approach to detect fallen people suitable for real cluttered scenes;
- its integration with two basic robot functionalities, 2D mapping and navigation, in order to suppress false

positives thanks to the a-priori knowledge of the environment and the availability of multiple view points;

- our RGB-D dataset of fallen people<sup>1</sup> consisting of several static and dynamic sequences with 15 different people acquired in 2 different environments.

The remainder of the paper is organized as follows. Section II reviews the work related to fall detection, people detection and body pose estimation. Section III describes our novel approach, first giving a picture of the entire workflow, then focusing on both the single-view approach and its integration with mapping and robot navigation. In Section IV, our dataset is described and our methods thoroughly evaluated. Finally, in Section V, conclusions are drawn and future directions of research identified.

## II. RELATED WORK

Nowadays the wide adoption of Deep Neural Networks (DNN) is boosting the classification accuracy in many fields. In particular, many recent works [8], [9], [10] address the person detection and body pose estimation problems showing great results. This kind of algorithms could be used also for detecting people lying on the floor. Nevertheless, their recognition capabilities are limited to RGB images and so they cannot work in dimmer scenes, which are usual in real life houses. In addition, the high complexity of the DNN requires the algorithm to be accelerated by using high-end Graphical Processing Units (GPU) in order to achieve real-time performances useful in real applications. For these reasons, those networks do not fit our application. Indeed, we are proposing techniques which can work also without the presence of the color information (e.g. under different illumination conditions or during the night). Moreover, we want to keep the power requirements at a minimum, given that this is a major issue in the design of mobile robots. Thus, the usage of an high-end GPU is unsuitable. Our approach draws upon two recent methods for the semantic segmentation of scene structures and objects from RGB-D data [11], [12]. Both approaches are almost real-time and based on fast features calculated on 3D patches or clusters. They also try to learn contextual relations among them, respectively by means of Conditional Random Fields and 3D Entangled Forests.

There exist also more specific approaches addressing the detection of falls. These comprehend wearable devices, whose great popularity is linked to the spread of open-source platforms which are small, powerful and connectible to low-cost sensors [13]. In most cases, such sensors include

The authors are with the Intelligent Autonomous Systems Laboratory (IAS-Lab), Department of Information Engineering (DEI), University of Padova, Via Ognissanti 72, 35129, Padova, Italy. morris.antonello, marco.carraro, emg@dei.unipd.it

<sup>1</sup><http://robotics.dei.unipd.it/117-fall>

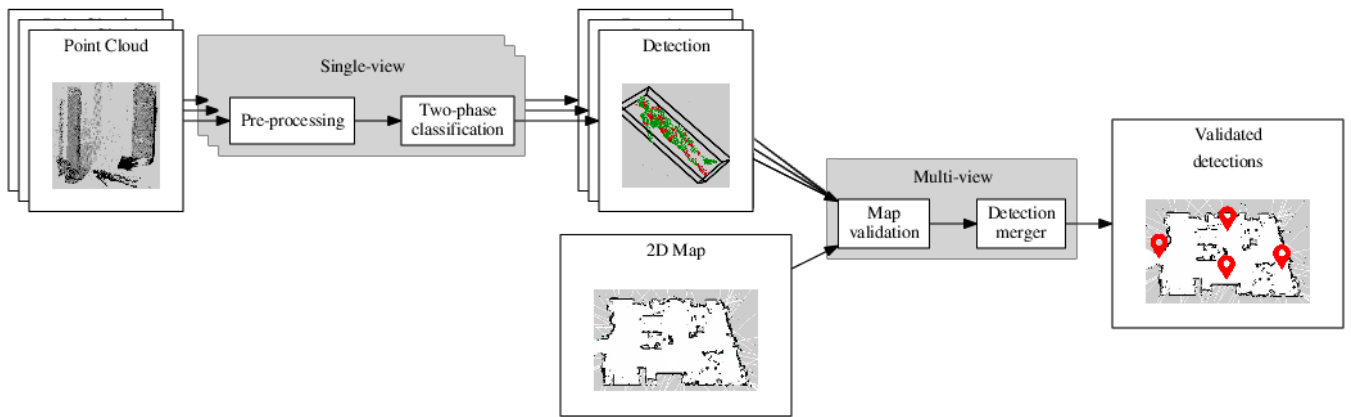


Fig. 1: The proposed approach is split into two separately running processes. The *single-view detector* detects fallen people on the single frames in a way which proves to be fast and robust to clutter. The *multi-view analyser* fuses the single-view results exploiting the availability of the 2D map and the multiple points of view explorable during the robot navigation. The final map includes also the semantic information about the location of the fallen people, see the red placeholders.

accelerometers [14], [15], [16]. These technologies suffer from the difficulty of correctly distinguishing falls from common actions like sitting or lying down. Furthermore, the elderly easily forget to wear them. Other approaches specifically addressing falls need the installation of environmental devices like microphones [17], cameras for person tracking [18], [19], [20], infrared or vibration sensors [21]. Anyway, these approaches are less effective and, being invasive, less accepted.

To the best of our knowledge, there exist just a few approaches trying to detect fallen people already lying on the floor: [22], [23], [24]. Both [22] and [23] are specifically designed for mobile robots. In [22], the authors propose a pipeline working on just single RGB images extending a deformable part-based model to the multi-view case for viewpoint invariant lying posture detection. Like us, [23] proposes a pipeline working on single depth images. Putative candidates are found by means of a segmentation phase based on an Euclidean clustering. Then, they are layered so as to face with occlusions and classified by means of a SVM using Histograms of Local Surface Normals. The downside of the approach is the Euclidean segmentation, in particular its distance threshold: if people fall on or near furniture, the segmented object may contain the user and parts of the furniture. On the contrary, this work specifically addresses this problem by concatenating two classifiers. Unfortunately, neither the code or dataset of [23] are available making a direct comparison impossible. Finally, in [24], a method for detecting and locating the head of a person lying on the floor by means of a RGB-D sensor is proposed. It would allow to test vital signs on the fallen people, but has not been tested in real cluttered scenarios and requires the head to be visible. Remarkably, none of the previous approaches take advantage of the other functionalities available thanks to the mobile robot like 2D mapping, i.e. the actual knowledge of the environment, and navigation, i.e. the availability of multiple view points.

### III. APPROACH

An overview of the proposed approach for detecting people lying on the floor is given in Figure 1. It is decoupled into two separately running processes, the *single-view detector* and the *multi-view analyser*. The former process, the *single-view detector*, operates on pure-3D Point Clouds generated by a RGB-D sensor such as the Kinect One V2, which, in our experiments, is mounted on a mobile robot 1.16 m off the floor and parallel to it. First, the input cloud is preprocessed in order to restrict the subsequent phases to work on a region of interest comprehending all the objects above the floor and below a maximum height. Then, the pre-processed cloud is over-segmented into small patches of voxels with similar appearance. In a two-phase classification step, the patches are classified as part of person or not and gathered together. The use of the Euclidean clustering on the cloud including only the person patches makes it possible to handle also cluttered scenes. Finally, to further improve performances, the latter process, the *multi-view analyser*, rejects all the detections not belonging to the free space of the 2D map and accumulates the detections from several frames by taking into account their 2D map positions and timestamps. Each phase is deeply discussed in the next subsections: Subsection III-A deals with the description of the *single-view detector* while Subsection III-B and Subsection III-C describe the *multi-view analyser*.

#### A. Patch-based Detection of Fallen People

Each point cloud is pre-processed to restrict the analysis to a region of interest and reduce the data noise. First of all, the point cloud is truncated to a 3D region containing the floor and the points between it and a maximum height of 0.66 m. Then, the floor is removed with an approach based on the RANSAC segmentation [25]. To improve its robustness to the robot motion, two floor planes are estimated, on a first half of the cloud close to the robot and on a second half far from the robot. In particular, a good split distance proved to be 3 m. Finally, to reduce the data noise without affecting

the running time, a soft statistical outlier removal is applied with the number of neighbours set to 50 and the standard deviation set to 0.3.

The core of the algorithm draws upon two recent works about the semantic segmentation of objects and scene structures [11], [12]. It comprehends the following 4 phases:

- 1) supervoxel over-segmentation in 3D patches;
- 2) classification of each 3D patch as positive, i.e. part of a fallen person or negative, i.e. not part of a fallen person;
- 3) clustering of positive patches;
- 4) classification of each cluster as positive, i.e. a fallen person, or negative, i.e. not a fallen person.

They allow to segment and classify correctly also the people lying close to other objects or scene structures. In the following, each phase is described.

The pre-processed point cloud is over-segmented into homogeneous 3D patches by means of the Voxel Cloud Connectivity Segmentation (VCCS) [26]. An example of over-segmented cloud is reported in Figure 2. This solution preserves the edges by finding patches not crossing object boundaries and, at the same time, it reduces the noise and the amount of data. The set of parameters used here is: voxel resolution 0.06 m, seed resolution 0.12 m, color importance 0.0, spatial importance 1.0 and normal importance 4.0. The voxel resolution is a good trade off between speed and having a sufficient number of points per patch. The seed resolution is a good trade off between having big patches and over-segmenting also the thinner body elements, e.g. arms and legs. The others are suggested in [27]. As the proposed approach does not rely on RGB data, color is not considered at all by setting the color importance to 0.0.

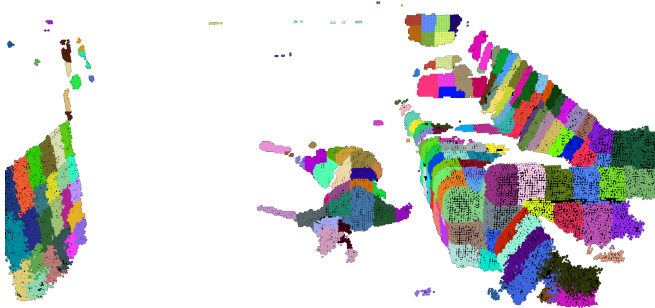


Fig. 2: An example of pre-filtered and over-segmented cloud. A random color is assigned to each patch. The person is lying in the center.

For each patch generated by the over-segmentation, a feature vector  $\mathbf{x}_1$  of length 16 is calculated. The choice of the features is based on the semantic segmentation works [11], [12], whose presented features proved to be as fast as effective. Here, the color features are left out and only the geometric features are taken into account. Some of them are calculated from the eigenvalues of the scatter matrix of the patch,  $\lambda_0 \leq \lambda_1 \leq \lambda_2$  while others from the Oriented

Bounding Box (OBB) including all the patch points. The complete list is given in Table I. To calculate the predicted

TABLE I: List of features calculated for each 3D patch and their dimensionality.

Features	Dimensionality
Compactness ( $\lambda_0$ )	1
Planarity ( $\lambda_1 - \lambda_0$ )	1
Linearity ( $\lambda_2 - \lambda_1$ )	1
Angle with floor plane (mean and std. dev.)	2
Height (top, centroid, and bottom point)	2
OBB dimensions (width, height and depth)	3
OBB face areas (frontal, lateral and upper)	3
OBB elongations ( $\frac{height}{width}, \frac{depth}{width}, \frac{height}{depth}$ )	3
Total number of features	16

label (part or not part of a fallen person) for each patch, this feature vector is then passed to a binary SVM classifier. After k-fold validation, a Radial Basis Function (RBF) kernel with the misclassification cost  $C$  equal to 62.5 and the bandwidth  $\gamma$  equal to 0.51 turned out to be the best performing solution. Of course, having each patch classified as part of a person (positive) or not (negative) does not suffice to detect a fallen person. Indeed, as shown in Figure 3(a), given that this classifier analyses just small patches, there can be false positives and false negatives. Because of this, two further steps, explained in the following and sketched in Figure 3(b)(c), have been developed in order to find 3D regions with a high density of positive patches and whose size is comparable to that of a person.

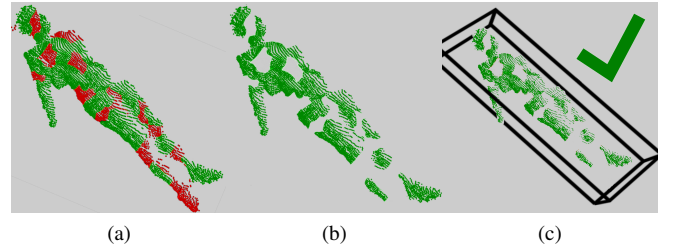


Fig. 3: The last three steps of the algorithm core: a) The first SVM classifies each patch as a person part (green color) or not (red color); b) Euclidean clustering of the positive patches; c) Calculation of the cluster OBB. The second SVM classifies each cluster as a person or not. Here, the response is positive.

In contrast to the methods in [23], having two sets of patches respectively with the positive and negative ones opens up the possibility to apply the Euclidean cluster extraction without the risk of segmenting a fallen person together with the adjacent scene elements. First of all, some false positive patches can be easily recognized, e.g. all the patches with less than 5 neighbouring positive patches in a radius of 0.5 m can be filtered out. Then, the negative patches are pushed aside, and the Euclidean clusters are extracted from the point cloud of the remaining positive patch centroids using a large distance threshold of 1.0 m.

For each cluster, its OBB is calculated. Thus, depending on the OBB dimensions and the number of positive and negative patches in it, each cluster may be a fallen person or not. For each cluster, a feature vector  $\mathbf{x}_2$  of size 9 has been devised. The complete list of features is given in Table II. In particular, the sample distances to the separating hyperplane returned by the former SVM turned out to be really useful. They have been exploited by means of an histogram with 4 bins for the distance intervals  $[0, 0.25)$ ,  $[0.25, 0.5)$ ,  $[0.5, 1)$  and  $[1, \infty)$ . For each cluster, each histogram bin is filled with the positive patches whose distance to the hyperplane falls in the respective interval. Thus, the number of positive patches in each bin/interval gives 4 additional features. The whole feature vector is passed to a binary SVM classifier. After k-fold validation, a RBF kernel with the misclassification cost  $C$  equal to 312.5 and the bandwidth  $\gamma$  equal to  $2.25 \times 10^{-3}$  turned out to be the best performing solution.

TABLE II: List of features calculated for each 3D cluster and their dimensionality.

Features	Dimensionality
OBB dimensions (width, height and depth)	3
Number of positive patches	1
Percentage of positive patches	1
4-bin histogram of positive patch confidences	4
Total number of features	9

### B. Map Verification

A mobile robot navigates through the environment thanks to the information of two maps: a static one necessary to compute a collision-free plan with static objects, e.g. walls or furniture items, and a dynamic one necessary to avoid moving obstacles, e.g. people. In this work, the static map, which is usually acquired only once and for all, is exploited to implement a false positive rejection phase. Let the static map be defined as a set of cells  $S = \{Cell_i, 0 \leq i \leq N\}$ , where:

$$Cell_i = \begin{cases} -1 & \text{unknown content} \\ 0 & \text{free space} \\ 0 < n \leq 1 & \text{probability to be occupied} \end{cases} \quad (1)$$

Thanks to the transformations computable with a 2D SLAM algorithm like [28], [29], each single-view detection can be transformed from the camera coordinate system to the map coordinate system and projected to a cell map  $Cell_i$ . If the  $Cell_i$  value is unknown ( $-1$ ) or occupied by a static obstacle ( $K \leq Cell_i \leq 1$  with  $K = 0.30$ ), then the detection can be easily rejected. An example of successful false positive rejection is shown in Figure 4, in which a single-view detections falls on the static furniture, in this case a tree trunk. Indeed, given its geometric similarity to a lying person, the single-view algorithm may detect it as a person. The map verification allows to reject it, enhancing the final detection performances. This step handles also other challenging situations, like shelf glass surfaces which can be really noisy.



Fig. 4: An example of successful false positive rejection performed by the map verification step: a) shows the furniture item raising some false positives, a tree trunk (in green) b) shows that these detections (the blue squares on the right) are located onto the map occupied space.

### C. Merging Detections from Multiple Vantage Points

The map is not the only robot feature that can enhance the detection performances. Indeed, in a typical scenario, the robot is patrolling a known environment. Thus, given that the location of each fallen person is mostly static, all the single-view detections available from the multiple points of view can be easily tracked. A detection may be a false positive from a certain view, while a true negative from many others. Moreover, the false positive detection rate is very low compared to the true positive one. Given these two facts, another contribution of this work is the exploitation of the detections available from the different vantage points.

After the map verification, the single-view detections are already expressed in the map reference system. In this section, an algorithm able to cluster or reject each of them is devised. Its output is a set  $\mathfrak{P}$  of validated lying person locations  $p_i$  in the map, formally  $\mathfrak{P} = \{p_i, 0 \leq i \leq g\}$ , where  $g$  is the total number of people. Given each new detection  $d = (loc, t)$ , where  $d.loc$  is its location in the map coordinate system and  $d.t$  its timestamp, the set of clusters, formally  $\mathfrak{C} = \{C_i : 0 \leq i \leq n\}$ , is updated with the following rule:

$$C_i = \{d_j : ||d_j.loc - d_m.loc|| < \overline{th}, \forall j, m \in [0, k-1]\}, \quad (2)$$

in which  $\overline{th}$  is a user-defined threshold which indicates if a detection is close enough to be considered in the cluster or not, and  $k$  the number of detections in the cluster.

The set  $\mathfrak{P}$  of fallen people is computed by a fixed-time periodic thread which analyses the set  $\mathfrak{C}$ . It updates the set  $\mathfrak{P}$  by deleting the old detections and analysing the new ones in  $\mathfrak{C}$ . Indeed, in order to maintain a lightweight representation of  $\mathfrak{C}$  and reject the false positives, whose frame rate is typically low, the old detections are discarded and a further check on the timestamp is performed. The pseudo-code of the whole procedure is reported in Algorithm 1, in which  $\hat{f}$  is the minimum detection frequency,  $\hat{t}$  is the maximum detection age and  $\hat{n}$  is the minimum number of detections in a cluster. Lines 3-7 handle the time-based rejection on the basis of the maximum allowed age, while Lines 8-14 reject the clusters whose detections have a low frame rate or are less than the minimum allowed.



In our implementation, we used  $\overline{th}$  equal to 1 m,  $\hat{t}$  equal to 60 s,  $\hat{f}$  equal to 1 Hz and  $\hat{n}$  equal to 5. The use of the frame rate allows to set a low  $\hat{n}$ , thus preventing over-fitting. The procedure is invoked by the periodic thread every 10 seconds. In Figure 5, the algorithm is shown in action.

**Algorithm 1** Cluster validation for detecting fallen people exploiting multiple vantage points

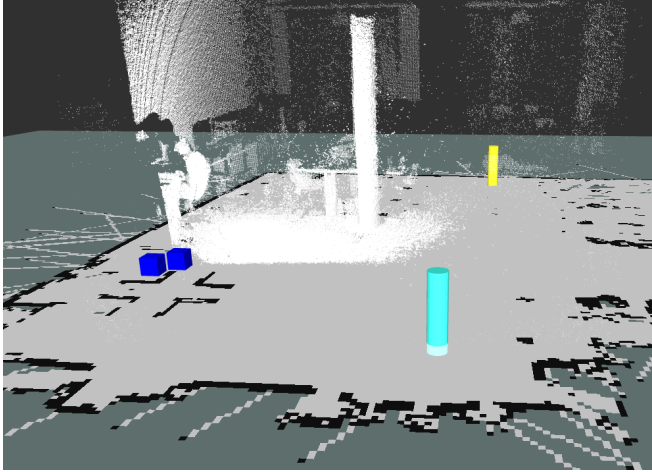
---

```

1: procedure VALIDATE_CLUSTERS( $\mathcal{C}, \mathfrak{P}, \hat{t}, \hat{f}, \hat{n}$ )
2:   for each  $C_i \in \mathcal{C}$  do
3:     for  $j \in [0, k-2]$  do
4:       for  $o \in [j+1, k-1]$  do
5:         if  $|d_o.t - d_j.t| > \hat{t}$  then
6:            $index \leftarrow \text{ARG\_MIN}(d_o.t, d_j.t)$ 
7:            $C_i \leftarrow C_i \setminus d_{index}$ 
8:        $t_m \leftarrow \min_{d \in C_i} \{d.t\}$ 
9:        $t_M \leftarrow \max_{d \in C_i} \{d.t\}$ 
10:       $f_i \leftarrow \frac{\|\mathcal{C}\|}{t_M - t_m}$ 
11:      if  $f_i \geq \hat{f}$  and  $\|C_i\| \geq \hat{n}$  then
12:         $loc_i \leftarrow \sum_{d \in C_i} \frac{d.loc}{\|C_i\|}$ 
13:         $\mathfrak{P} \leftarrow \mathfrak{P} \cup loc_i$ 
14:         $\mathcal{C} \leftarrow \mathcal{C} \setminus C_i$ 
15:  return  $\mathcal{C}, \mathfrak{P}$ 

```

---



(a)

Fig. 5: The single-view detections projected on the 2D map are analysed by the *multi-view analyser*. If they meet both the distance and time criteria, they are clustered. The white points compose the input point cloud, the blue cubes are the projected detections, here rejected false positives, and the coloured cylinders are the validated detection.

#### IV. RESULTS

The detection of fallen people is a challenging problem also because of the lack of public datasets. For this reason, another contribution of this work is the release of the IASLAB-RGBD Fallen Person Dataset<sup>2</sup>. On it, 4 common

metrics, the detection *accuracy*, *precision*, *recall* and  $F_{0.5}$  score, are evaluated for each presented method. If  $TP$ ,  $TN$ ,  $FP$  and  $FN$  are the true positives, true negatives, false positives and false negatives, then these metrics are defined as in the following:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F_{0.5} = \frac{(1 + 0.5^2) * precision * recall}{0.5^2 * precision + recall}, \quad (6)$$

where the  $F_{0.5}$  score, already proposed in [23], is an harmonic average of precision and recall promoting an high precision, i.e. a low number of false positives. In addition, given the impossibility to compare with other existent and similar approaches, the baseline to which our algorithms are compared is a simple approach based on the Euclidean cluster extraction. This way, it will be clear how important the use of patches is in order to handle cluttered scenes. Finally, a detailed analysis of the running times is provided.

##### A. IASLAB-RGBD Fallen Person Dataset

This dataset consists of several RGB-D frame sequences containing 15 different people. It has been acquired in two different laboratory environments, the *Lab A* and the *Lab B*, by means of a Microsoft Kinect One V2, placed on a pedestal or on our mobile robot. The *Lab A* is bigger and useful to test whether the algorithm can find people in the full distance range of the sensor (up to 5 m). The *Lab B* is smaller and more similar to a real domestic scenario. It is more cluttered and contains a sofa. It comprehends also glass surfaces which can be very noisy. For the sake of explanation, the dataset can be divided into three parts:

- 1) Part 1 includes 360 RGB-D frames acquired from 3 static pedestals. It is composed of several views of 10 people, which have been asked to lie in 12 different poses, 6 from the back and 6 from the front. Each person has been manually segmented in 3D;
- 2) Part 2 includes 4 sequences of RGB-D frames, for a total of 15932 frames, acquired from a mobile robot during its patrolling task in the *Lab A*. People lie in 4 different fixed locations;
- 3) Part 3 includes 4 sequences of RGB-D frames, for a total of 9391 frames, acquired from a mobile robot during its patrolling task in the *Lab B*. People lie in 4 different fixed locations.

Training and test splits are also available. Some images of the dataset will be reported when discussing the results even if our approach does not exploit the RGB info.

The first classifier of the *single-view detector* has been trained on thousand of patches extracted from the frames in Part 1 and Part 2 and tested on patches extracted from the

<sup>2</sup><http://robotics.dei.unipd.it/117-fall>

frames in Part 1 and 3. All the positive samples have been taken from Part 1. The 70-30 train-test split of the segmented fallen people in Part 1 is also available. Negative samples have been taken from the *Lab A* (just 24 frames out of 15932), the *Lab B* (just 32 frames out of 9391) and the NYU Depth Dataset V2 [30] (just 35 out of 1449), which contains thousands of indoor scenes for scene understanding. Only some of the negative samples have been used for balancing the number of positive and negative samples.

The second classifier of the *single-view detector* has been trained on clusters extracted from the frames in Part 2 (*Lab A*) and tested on clusters extracted from the frames in Part 3 (*Lab B*). Approximately, for the training, the 15% of all the available frames has been considered.

Not only the *single-view detector* but also the *multi-view analyser* has been tested on Part 3. Indeed, both Part 2 and 3 comprehend the entire robot transformation tree. Given that the position of the fallen people in the 2D map is known, this allows to calculate the performance indices automatically by checking if the location of the detected cluster centroid is close (at a distance less or equal to 1 m) to the ground truth centroid of a person position in the 2D map.

### B. Validation

The presented methods have been quantitatively evaluated on the IASLAB-RGBD Fallen Person Dataset. First, the separated evaluation of each classifier is presented. Then, the entire pipeline has been evaluated on both rooms, the *Lab A* and the *Lab B*. As previously explained, both the classifiers have been trained on just a part of the frames in the *Lab A* while they see the *Lab B* for the first time. In particular, we present the results for each of the 3 contributions, the *single-view detector* and the two modules of the *multi-view analyser*: the map validation and the detection merging from multiple vantage points. Furthermore, given the impossibility to compare directly with [23], the comparison baseline (B) is a simple approach not exploiting patches. It finds putative clusters by means of the Euclidean cluster extraction with a distance threshold of 0.10 m, which is really low considering a voxel resolution of 0.06 m and far less than the one required by our approach (1 m). The baseline classifies then each cluster on the basis of its position and its OBB size.

As previously mentioned, both classifiers of the *single view detector* have been trained and tested on two different dataset splits. In both cases, K-fold validation with  $K$  equal to 10 has been performed on the training set in order to find the optimal misclassification cost  $C$  and bandwidth  $\gamma$  values of the RBF kernel. As a preliminary evaluation, the SVM performances on the respective test sets are reported in Table III.

TABLE III: Performances of the two classifiers on their test sets.

Method	Accuracy	Precision	Recall	$F_{0.5}$
Classifier 1 (C1)	0.89	0.93	0.84	0.91
Classifier 2 (C2)	0.93	0.86	0.95	0.88

The results of the quantitative comparison between all the methods are shown in Table IV and V. Thanks to the patches, our methods outperform the baseline, not only in precision but also in recall. Furthermore, the map validation can further improve performances by rejecting some false positives.

TABLE IV: Performance comparison on the *Lab A*.

Method	Accuracy	Precision	Recall	$F_{0.5}$
Baseline (B)	0.88	0.65	0.33	0.54
Single-view (SV)	0.90	0.77	0.78	0.77
SV + Map verification (MV)	0.92	0.87	0.77	0.85

TABLE V: Performance comparison on the *Lab B*, never seen before by both classifiers.

Method	Accuracy	Precision	Recall	$F_{0.5}$
Baseline (B)	0.84	0.64	0.26	0.50
Single-view (SV)	0.89	0.87	0.74	0.83
SV + Map verification (MV)	0.90	0.92	0.72	0.87

As shown in Table VI, also the detection merging from multiple vantage points proved to be useful. It has been tested on each one of the eight frame sequences acquired in the two environments. Each time, even if the environment is the same, the navigation path can differ due to dynamic obstacles and the different positions of the lying people on the floor. After the 4 patrolling tasks of the *Lab A*, each person is always detected and only once, a false positive is still present while, after the 4 patrolling tasks of the *Lab B* (never seen before by both classifiers), each person is always detected and all the false positives are successfully rejected.

TABLE VI: Performances of the *multi-view analyser* on both environments. Each time, even if the environment is almost the same, the robot path can differ because of dynamic obstacles and different positions of the lying people on the floor.

Environment	TP/P	FP
Lab A (sequence 1)	4/4	0
Lab A (sequence 2)	4/4	1
Lab A (sequence 3)	4/4	0
Lab A (sequence 4)	4/4	0
Lab B (sequence 1)	4/4	0
Lab B (sequence 2)	4/4	0
Lab B (sequence 3)	4/4	0
Lab B (sequence 4)	4/4	0

Finally, in Figure 6, some qualitative results are reported. They show the ability of the *single-view detector* to find people in cluttered environments, see Figure 6(a)(b)(c)(d). Two difficult cases due to close objects or noisy regions, like glass surfaces, are also reported, see 6(e)(f). Anyway, they are easily handled by the *multi-view detector*, 6(g)(h).

### C. Runtime Analysis

In Table VII, the running times of *single-view detector* are reported. The algorithm is very efficient in terms of



Fig. 6: Qualitative results on the IASLAB-RGBD Fallen Person Dataset: (a)(b) even if the lying people can be very close to the wall or other scene elements, the *single-view detector* can find them at a high detection rate; (c)(d) the *single-view detector* can discard fake lying people, see the white circles; (e)(f) the *single-view detector* may find some false positives in the presence of clutter (several close objects) or high noise (glass surfaces); (g)(h) the *multi-view analyser* can reject both FP like in (e) thanks to the low frame rate or in (f) thanks to the map validation.

computing time proving to be an optimal choice for a mobile robot. Even if it is not yet fully parallelized, it can work in real-time at an average speed of 7.72 fps. The test machine is a Dell Inspiron 15 7000 with an Intel Core i7-6700HQ CPU with 4 cores clocked at 2.60GHz, 16 GB of RAM and Linux Mint 17.3. Given that the *multi-view analyser* is a daemon running in the background, its running times are of no interest and thereby not reported.

TABLE VII: Average runtimes of the main steps of the proposed algorithm on our test machine (Intel Core i7-6700HQ CPU, 2.60GHz x 4).

Processing Stage	Runtime
Pre-processing and Oversegmentation	10.27 fps
Patch Feature Extraction	105.98 fps
SVM Classification 1 (per patch)	0.84 $\mu$ s
Cluster Feature Extraction	2639.56 fps
SVM Classification 2 (per cluster)	0.04 $\mu$ s
Total runtime	7.72 fps

## V. CONCLUSIONS

This paper presented a real-time and robust approach to detect fallen people lying on the floor in various positions and from different distances. A single-view algorithm, which draws upon recent developments in the semantic segmentation field and does not need restrictive distance thresholds to segment putative clusters, was fully integrated on a mobile robot. The map of the environment and the availability of many different vantage points allowed to reduce the number of false positives, further improving the final performances. The algorithms here presented were thoroughly validated on the IASLAB-RGBD Fallen Person Dataset, which was published online for the benefit of the research community. They clearly outperform a simple method based on a finer distance threshold. In the near future, we would like to validate not only the ability of the algorithm to detect, but also to semantically segment fallen people. We also plan to extend the test bed with sequences taken from real apartments along with different navigation paths. Finally, it would be interesting to merge close similar patches before

their classification in order to analyse bigger segments.

## ACKNOWLEDGEMENT

The research leading to these results has been partially supported by Omitech Srl. The authors would like to thank M. Munaro and S. Ghidoni for valuable discussions.

## REFERENCES

- [1] W. He, D. Goodkind, and P. Kowal, "An aging world: 2015," tech. rep., 2016.
- [2] M. E. Pollack, L. Brown, D. Colbry, C. Orosz, B. Peintner, S. Ramakrishnan, S. Engberg, J. T. Matthews, J. Dunbar-Jacob, C. E. McCarthy, et al., "Pearl: A mobile robotic assistant for the elderly," in *AAAI workshop on automation as eldercare*, vol. 2002, pp. 85–91, 2002.
- [3] F. Cavallo, M. Aquilano, M. Bonaccorsi, R. Limosani, A. Manzi, M. C. Carrozza, and P. Dario, "On the design, development and experimentation of the astro assistive robot integrated in smart environments," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pp. 4310–4315, IEEE, 2013.
- [4] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter, and N. Doering, "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 5992–5999, IEEE, 2015.
- [5] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, et al., "Hobbit, a care robot supporting independent living at home: First prototype and lessons learned," *Robotics and Autonomous Systems*, vol. 75, pp. 60–78, 2016.
- [6] M. Carraro, M. Antonello, L. Tonin, and E. Menegatti, "An open source robotic platform for ambient assisted living," *Artificial Intelligence and Robotics (AIRO)*, 2015.
- [7] S. Lord, C. Sherrington, H. Menz, and J. Close, *Falls in older people: risk factors and strategies for prevention*. Cambridge University Press, 2007.
- [8] A. Angelova, A. Krizhevsky, V. Vanhoucke, A. S. Ogale, and D. Ferguson, "Real-time pedestrian detection with deep network cascades," in *BMVC*, pp. 32–1, 2015.
- [9] T. Cao, Z. and Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.
- [10] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4732, 2016.
- [11] D. Wolf, J. Prankl, and M. Vincze, "Fast semantic segmentation of 3d point clouds using a dense crf with learned parameters," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 4867–4873, IEEE, 2015.
- [12] D. Wolf, J. Prankl, and M. Vincze, "Enhancing semantic segmentation for robotics: the power of 3-d entangled forests," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 49–56, 2016.
- [13] J. Perry, S. Kellog, S. Vaidya, J.-H. Youn, H. Ali, and H. Sharif, "Survey and evaluation of real-time fall detection approaches," in *High-Capacity Optical Networks and Enabling Technologies (HONET), 2009 6th International Symposium on*, pp. 158–164, IEEE, 2009.
- [14] Q. Li, J. Stankovic, M. Hanson, A. Barth, J. Lach, and G. Zhou, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in *Wearable and Implantable Body Sensor Networks, 2009. BSN 2009. Sixth International Workshop on*, pp. 138–143, IEEE, 2009.
- [15] J. Boyle and M. Karunanithi, "Simulated fall detection via accelerometers," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 1274–1277, IEEE, 2008.
- [16] U. Lindemann, A. Hock, M. Stuber, W. Keck, and C. Becker, "Evaluation of a fall detector based on accelerometers: A pilot study," *Medical and Biological Engineering and Computing*, vol. 43, no. 5, pp. 548–551, 2005.
- [17] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 4628–4631, IEEE, 2008.
- [18] A. Williams, D. Ganesan, and A. Hanson, "Aging in place: fall detection and localization in a distributed smart camera network," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 892–901, ACM, 2007.
- [19] R. Cucchiara, A. Prati, and R. Vezzani, "A multi-camera vision system for fall detection and alarm generation," *Expert Systems*, vol. 24, no. 5, pp. 334–345, 2007.
- [20] S. Ghidoni, S. M. Anzalone, M. Munaro, S. Michieletto, and E. Menegatti, "A distributed perception infrastructure for robot assisted living," *Robotics and Autonomous Systems*, vol. 62, no. 9, pp. 1316–1328, 2014.
- [21] A. Yazar, F. Erden, and A. E. Cetin, "Multi-sensor ambient assisted living system for fall detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 14)*, pp. 1–3, Citeseer, 2014.
- [22] S. Wang, S. Zahir, and B. Leibe, "Lying pose recognition for elderly fall detection," *Robotics: Science and Systems VII*, vol. 345, 2012.
- [23] M. Volkhardt, F. Schneemann, and H.-M. Gross, "Fallen person detection for mobile robots using 3d depth data," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pp. 3573–3578, IEEE, 2013.
- [24] K. Nishi and J. Miura, "A head position estimation method for a variety of recumbent positions for a care robot," in *Proceedings of the 6th Int. Conf. on Advanced Mechatronics*, 2015.
- [25] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [26] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel cloud connectivity segmentation-supervoxels for point clouds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2027–2034, 2013.
- [27] S. C. Stein, F. Wörgötter, M. Schoeler, J. Papon, and T. Kulvicius, "Convexity based object partitioning for robot applications," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 3213–3220, IEEE, 2014.
- [28] G. Grisetti, C. Stachniss, and W. Burgard, "Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 2432–2437, IEEE, 2005.
- [29] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE transactions on Robotics*, vol. 23, no. 1, pp. 34–46, 2007.
- [30] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, pp. 746–760, Springer, 2012.